BIOLOGICAL DATA INTEGRATION USING VIRTUAL DATABASE

Ateet Mehta, Dr. Kalpesh Lad, Dr. Bankim Patel

ABSTRACT

Biological data integration is considered to be one the most important and challenging tasks in bioinformatics. The scientific achievements greatly depend on the integrated view of largely diverse set of data. Biological data reside in hundreds of database and there is no single database providing an integrated view of data. It greatly invokes the need of data integration. Though they are different approaches for data integration like data warehouse, federation, webservices; each has its own pros and cons and challenges of implementation. In this research paper, we have proposed a framework using virtual database to integrate different biological data sources.

Keywords: Data Integration, Data Warehouse, Data Federation, Webservice, virtual database

1. INTRODUCTION

The scientific achievements coming from molecular biology greatly depends on integrated view of the data dispersed in different databases managed by different data source providers [1]. According to the last release of the Nucleic Acid Research, there are more than 1170 databases in the field of molecular biology [2]. Each database corresponds to the output of a specific study or community and represents a huge investment whose potential have not been fully explored. Until recently, data sources were set up as autonomous websites by individual institutions or research laboratories [3]. Data Sources vary considerably in contents, access methods, data formats, capacity, data access methods and services [4].

To understand how molecules, and ultimately cells, function in tissues, organs and organisms, biologists need to adapt a systems-level approach in biology [5]. In other words, the pendulum of bioscience, swinging towards the integrated view of biology, will depend critically on the integration from a variety of data sources. Over the past two decades, research in evolutionary biology has come to depend on sequence comparisons at the gene and protein level, and in the future, it will depend more and more on tracking not just DNA sequences but how entire genomes evolve over time [6-7]. It poses greater challenges firstly for

data integration and then efficient management of data [8]. It is widely recognized that successful data integration is one of the keys to improve productivity for stored data [9].

Though they are many approaches of data integration; each having their own advantages, limitations and challenges in implementation; we have proposed biological data integration using a virtual database. Next section discusses related work in the area followed by the solution that we have proposed.

2. RELATED WORK

Most biological data source providers have their own visual interfaces from which users can retrieve data. One approach is application linking in which the web pages from the source providers are linked in the external applications. However this approach is good for data visualization, it is not efficient for data analysis. Other approach is to parse the web pages of the source providers and integrate data in the local database for further analysis. This approach is better than application linking but has to rely greatly on source provider's page layouts and limits integrated view on data and capabilities of data integration. Data warehouse can be an effective solution to data integration however at the cost of maintaining data warehouse.

3. PROPOSED WORK

Our proposed model of data integration using virtual database provides capability to aggregate data from diverse databases into a virtual database. Unlike Data warehouse architecture in which data from diverse data sources are actually extracted, transformed and finally loaded into central database, the virtual database would store meta data about different data sources rather than storing actual data. So Virtual Database is basically a logical association of independent databases providing a single, integrated, coherent view of all resources logically registered in the virtual database. The architecture of virtual database looks like a single database to the end users and thereby hiding all the internal details and complexity from the user. In Bioinformatics, data sources are heterogeneous ranging from simple flat files, spreadsheets, CSV files, XML, HTML documents to ODBC compliant data bases.

Research institutes/Organizations can identify public data sources that provide data of their interest and register those data sources in their virtual database as a meta data. The meta data should contain location of the source database, access identifiers of source database, access methods, structure of data, relationships of the biological entity from one data sources with the biological entities provided

by other data sources and annotations required to formulate single query to access diverse data sources upon user's request.

The main characteristics of the virtual database are:

- Unified, and integrated view of data from multiple resources
- Logical integration of distributed datasets into wider view
- Data sources are diverse and can be public/private
- Data sources are heterogeneous in nature
- Data sources are geographically distributed

The Virtual database should contain, but not limited to, the following components.

- Query Execution Engine
- Query Transformation Component
- Data Transformation Component
- Meta Data Repository
- Collection of Wrappers each corresponding to distinct type of data source.

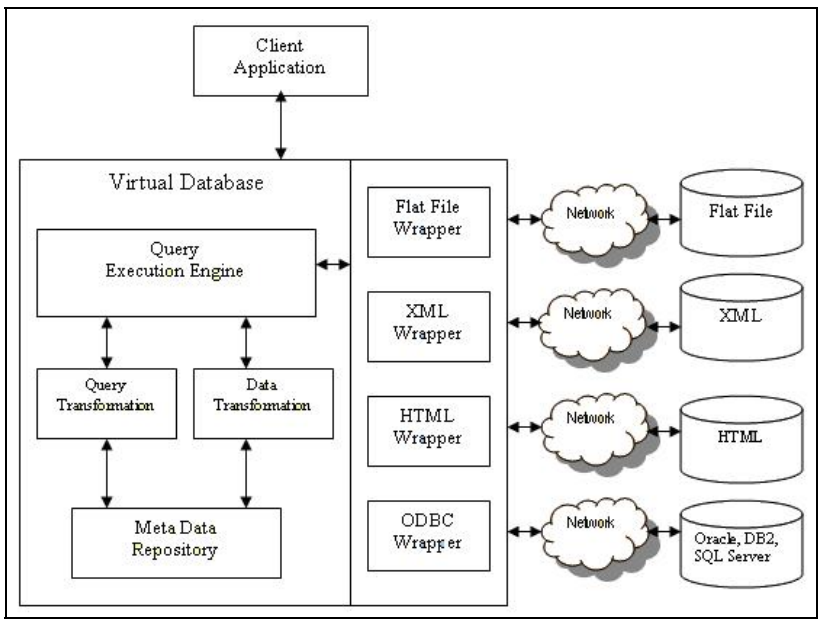

Fig. 1: Architecture of Virtual Database

Client interacts through API or Browser based tool with Virtual Database and passes query which in turn is formulated and executed by Virtual Database and the results are returned to the user. It hides all complexities from the user.

Next section discusses the components of the virtual database.

3.1 Query Execution Engine

Query Execution Engine is the key component of the virtual database which manages the overall execution of user's request to access data. It performs mainly following tasks.

- Receive request from client application to retrieve data
- Parse the request into different units to determine biological entity and data sources
- Call Query Transformation Unit to transform the single query into distinct queries for distinct data sources
- Distribute transformed query units to each of the wrappers which are involved to connect to particular type of data source
- Receive results from the wrappers
- Call Data Transformation Unit to aggregate data received from different wrapper to formulate a result set
- Return the results to the client

3.2 Query Transformation

The input to Query Transformation unit is the parsed query which it receives from the Query Execution Engine. The parsed query consists of set of query blocks which are interrelated to each other. The form of the query decides how the query blocks are interrelated to each other. The main job of the Query Transformation Unit is to formulate query for each distinct data sources using the query blocks it has received. So it will formulate one query into multiple queries. It will extract information from the meta data repository in order to determine the biological entity and the respective data source involved in the query.

For Example, the user might request following:

"Find structure of hypothetical protein HPF0109 in V. cholerae, homology models, classifications, and recent publication entries"

In this example, user has requested the three dimensional structure of a protein along with corresponding homology models, protein classification and related citations and publication information. Query Transformation Unit will recognize

the requested biological entity and corresponding data sources and formulates a single query into three distinct queries. Biological entities involved here are Protein, its homology model which can be searched in SWISS-PROT, Protein Classification which can be searched in PFAM and Publications/Citations which can be searched in PUBMED. The output of Query Transformation Unit is the logical forms of the query blocks distinctly made for data sources. Based on this transformation, the actual data retrieval request would be generated by the wrappers which are made for these data sources.

3.3 Wrappers

Data sources in bioinformatics are heterogeneous ranging from flat files, XML to ODBC compliant data sources and webservices. The access methods to these heterogeneous data sources are different; hence distinct wrappers are required to access these distinct data sources. Further, even for a same kind of a data sources such as ODBC Compliant data sources like Oracle, SQL Server, DB2, MS SQL etc, data source providers might have exposed different access mechanisms. The virtual data base will implement Wrapper for each distinct data source types considering data source type and access mechanism. The main job of a wrapper is to translate the logical query block it receives from Query Execution Engine in a more technical form which can be requested to the data source to retrieve data.

In case of ODBC compliant data sources, wrapper has to generate request in a form of SQL statements. The virtual database might contain in the meta data repository database links, view definitions, materialized views, synonyms for the database procedures and functions existing in the source database provided they are exposed by the data source providers. Wrapper will build SQL using meta data repository and send request to the data source providers. Upon receipt of the results, it delivers them back to the query execution engine.

Wrappers built for HTML data source will dynamically create URL based on the user's query and request the data source providers. HTML wrapper will have parser algorithms to parse HTML page to retrieve data from HTML page components and deliver it to Query Execution Engine. Similarly Wrappers for Flat file, CSV file and Spreadsheet will have parsing mechanism to retrieve data. Wrappers for XML will use XSD to parse data.

3.4 Data Transformation

The main job of Data Transformation Unit is to aggregate results from different wrappers and collates them in a single result set. Data Transformation also includes conversion of data types, transforming null values to blank space or special delimited characters, concertinaing more than one field into a single

field. It receives such format specific transformation details from the meta data repository.

3.5 Meta Data Repository

Meta Data is a repository within the Virtual database which stores information about different biological entities, relationships, corresponding data sources, data source access descriptors, database links, views, synonyms on object definitions in case of ODBC compliant database if they are exposed by the data source providers. The repository also contains transformation rules for different attributes of the biological entity. Meta Data can also optionally contain statistics information which query execution engine can use to optimize query.

4. CONCLUSION

Data integration using virtual database is an approach to integrate diverse set of data from different data source providers without the need to physically extract and load data into the local database. With the use of meta data, user's query is transformed for the target data source in more than one queries when needed. Similarly data sets from multiple data sources are transformed into a single data set as a result to the query. However, data integration in bioinformatics is at a very infant stage and more efficient data integration strategy, standards for data structures are needed to be established.

REFERENCES:

1. Critchlow Terence and Lacroix Zoe (2004) Bioinformatics- Managing scientific data, Morgan Kaufmann Publishers, San Francisco. pp. 1-441
2. M.Y. Galperin (2007), The Molecular Biology Database Collection: 2008 Update, Nucleic Acid Research.
3. C.A. Goble and R. Stevens (2001), Transparent Access to Multiple Bioinformatics Information Sources, IBM Systems Journal 40, Vol 2. pp. 532-552
4. W. Zhong and P.W. Sternberg (2007), Automated Data Integration for Developmental Biological Research, Development, Vol 133, pp. 3227-38
5. L.D. Stein (2003) Integrating biological databases, Nat Rev Genet, Vol 4, pp. 337-45
6. M.Y. Galperin (2007) The molecular biology database collection, Nucleic Acid Research
7. J. Arrais, B. Santos, J. Ferandes, L.Carreto, M.A.S. Santos and J.L. Oliveira (2007) GeneBrowser : an approach for integration and functional classification of genomic data, Journal of Integrative Bioinformatics, Vol 4
8. L. Wong (2002) Technologies for integrating biological data, Brief Bioinform, Vol 3, pp. 389-404

9. Joel Arrais, Joao E. Pereira, Joao Fernandes and Jose Luis Oliveira (2009) GeNs: a Biological Data Integration Platform, World Academy of Science, Engineering and Technology